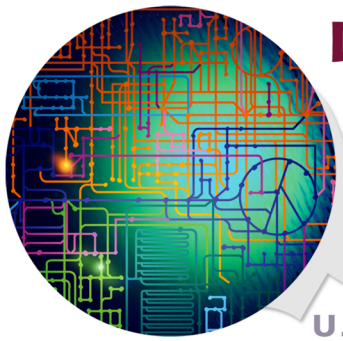


DOE GENOMICS:GTL

SYSTEMS BIOLOGY FOR ENERGY AND ENVIRONMENT

OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

This document

has been updated.

New file is available at

http://doegenomestolife.org/pubs.shtml

November 2005

DOE/SC-0069

Genomes to Life: Realizing the Potential of the Genome Revolution

Office of Science
January 2003
DOEGenomesToLife.org

The remarkable successes of the Human Genome Project and spin-offs revealing the details of numerous genomes—from microbes to plants to mice—provide the richest resource in the history of biology. These achievements now empower scientists to address the ultimate goal of modern biology: to obtain a fundamental, comprehensive, and systematic understanding of life. This goal is founded, as is life itself, on the genome, whose genes encode the proteins that carry out most cellular activities via a labyrinth of pathways and networks that make the cell "come alive" (see figure below).

Catalyzing Systems Biology

The Department of Energy's (DOE) Genomes to Life (GTL) program is combining high-throughput advanced technologies and computation with the information found in microbial genomes to establish a foundation for achieving an understanding of living systems (see "Microbes for DOE Missions," p. 2). GTL is designed to help launch biology onto a new trajectory to comprehensively understand cellular processes in a realistic context. This new level of exploration, known as systems biology, will empower scientists to pursue completely new approaches to discovery, transforming biology to a more quantitative and predictive science. GTL scientific goals target the fundamental processes of living systems by studying them on three levels:

1. Proteins and multicomponent molecular machines that form all of the cell's structures and perform most of the cell's work.
2. Gene regulatory networks and pathways that control cellular processes.
3. Microbial communities in which groups of cells carry out complex processes in nature.

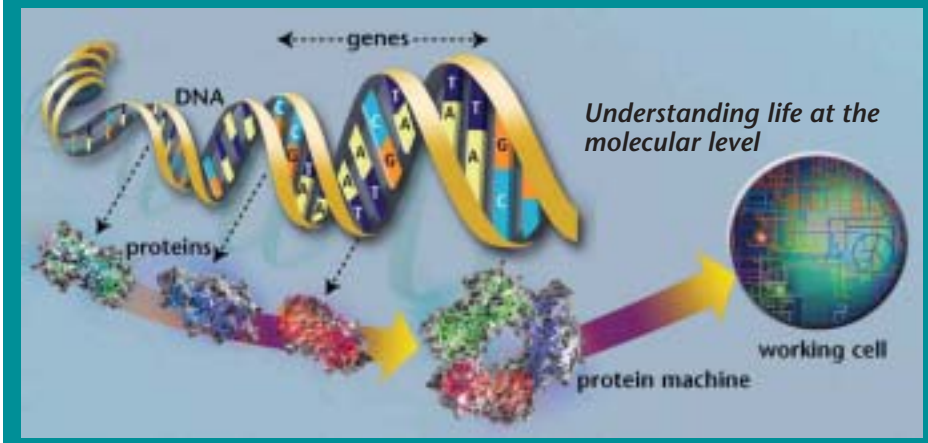
Molecular machines carry out chemical reactions, generate mechanical forces, transport metabolites and ions, and make possible every action of a biological system. A cell does not generate its entire repertoire of molecular machines at once. Genomic regulatory elements dictate the particular set produced according to the organism's life strategy and in response to environmental cues, including other microbial populations in the larger ecological community.

A comprehensive approach to understanding biological systems thus extends from individual cells to many cells functioning in communities. Such studies must encompass proteins, molecular machines, pathways, networks, cells, and, ultimately, their regulatory elements, cellular systems, and environments. This next generation of biology is viable only with vastly increased computational and informational capabilities to master the full complexities of biological systems.

Understanding life processes at the molecular level is a "national science priority."

—OSTP-OMB FY 2004 budget guidance memo; see p. 6.

Genomes to Life: From DNA Sequence to Living Systems



Genes are made up of DNA and contain the information used by other cellular components (e.g., RNA and ribosomes, not shown here) to create proteins. A working cell is tightly packed with tens of thousands of proteins and other molecules, often working together as multimolecular "machines" to perform essential cellular activities (see also cell figure, p. 5).

Transforming Biology

Just as DNA sequencing capability was completely inadequate at the beginning of the Human Genome Project (HGP), the quantity and complexity of data that must be collected and analyzed for systems biology research far exceed current capabilities and capacities. The HGP taught that aspects of biological research can be made high-throughput and cost-effective (see graph, p. 5). Collecting and using such data and reagents will require a new organizational model that coordinates and integrates dozens of high-throughput technologies and approaches, some not yet refined or even developed. This is the central principle of GTL and indeed of all systems biology research.

Analysis of living systems will require a new generation of experimentation and the computational methods and capabilities to assimilate, understand, and model the data on the scale and complexity of real living systems. Computing must guide the research questions and interpretation at every step.

The knowledge base resulting from the GTL program will provide the entire research community with data, models, and simulations of gene expression, pathways, and network systems; molecular machines; and cell and community processes. These new capabilities and resources will inspire revolutionary solutions to DOE mission needs and transform the entire life sciences landscape, from agriculture to human health.

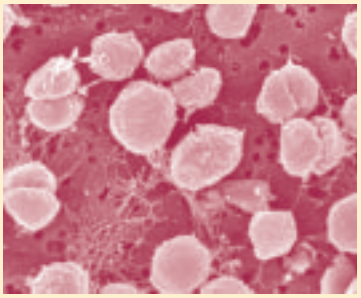
Microbes for DOE Missions: Energy Security, Cleanup, Climate Change

Why Study Microbes?

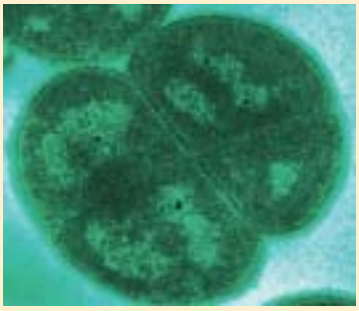
The ability of this planet to sustain life is largely dependent on microbes. They are the foundation of the biosphere, controlling earth's biogeochemical cycles and affecting the productivity of the soil, quality of water, and global climate. As one of the most exciting frontiers in biology today, microbial research is revealing the hidden architecture of life and the dynamic, life-sustaining processes on earth. The diversity and range of their environmental adaptations mean that microbes long ago "solved" many problems for which scientists are seeking solutions today (see examples at right). The incomprehensible number of microbes is an untapped but valuable resource that ultimately may be used to generate new energy sources (e.g., hydrogen for a new energy economy), new cleanup and industrial processes, and new ways of using biology to address DOE missions.

The Challenge

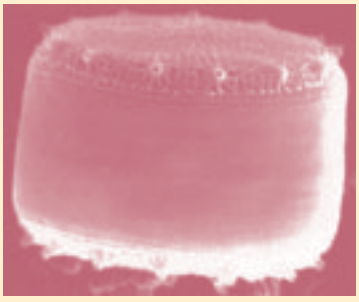
Microbes have become masters at living in almost every environment, harvesting energy in almost any form. Their sophisticated biochemical capabilities can be utilized for transforming wastes and organic matter, cycling nutrients, and, as part of the photosynthetic process, converting sunlight into energy and "fixing" (storing) CO_2 from the atmosphere. The analytical complexity involved in understanding these processes is enormous. Thousands of microbes have capabilities of interest. Moreover, each microbial genome contains thousands of genes capable of producing an even-greater number of proteins. These proteins combine to form innumerable molecular machines in myriad pathways and networks, many of which carry out biological processes useful for DOE missions (see "Potential Applications of GTL Science," p. 3).



Methanococcus jannaschii: Produces methane, an important energy source; contains enzymes that withstand high temperatures and pressures; possibly useful for industrial processes.



Deinococcus radiodurans: Survives extremely high levels of radiation and has high potential for radioactive waste cleanup.



Thalassiosira pseudonana: Ocean diatom that is major participant in biological pumping of carbon to ocean depths and has potential for mitigating global climate change.

Emerging Technologies and Computing for Systems Biology: Establishing a Firm Foundation for Genomes to Life

Numerous projects funded by the Office of Biological and Environmental Research (BER) over the past 5 years have established a strong foundation for the GTL program. These projects underscore the need for high-throughput biological research and novel computational approaches. They are also demonstrating the power of mass spectrometric analyses of whole microbial proteomes, developing new imaging methods, advancing the use of microarrays for expression analyses, exploring scalable ways to generate microbial proteins, and developing computational tools for second-generation genome analysis and annotation.

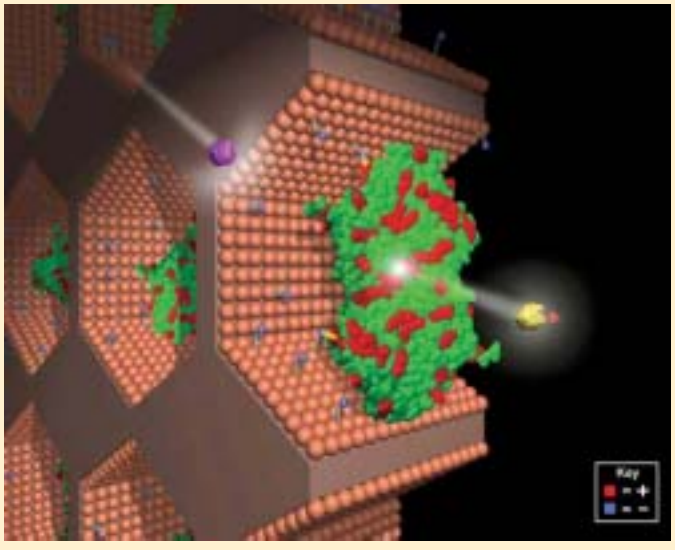
Several collaborative groups are integrating technologies and computational modeling to gain a systems understanding of specific microbes in their natural environments. For example, the *Shewanella* Federation, consisting of teams from academia, national laboratories, and other organizations is making progress in preliminary proteome and expression analyses of this remarkably versatile organism that can immobilize toxic uranium from ground water. By focusing multiple technologies on a single organism, the federation is integrating diverse experimental results into a multidimensional perspective of the biology of this key microbe. Thus far, the group has identified >77% of the predicted proteome of *Shewanella* (3782 of 4855 predicted genes) using ultrahigh-resolution mass spectrometry techniques. This and other groundbreaking BER projects (e.g., on *Deinococcus radiodurans*) have elucidated the highest percentages of the proteomes of organisms studied to date. These projects have also set the stage and identified the need for developing high-throughput user facilities accessible to the biological research community (see p. 5).

Office of Science—At the Forefront of the Biological Revolution

In 1986 the DOE Office of Science launched the Human Genome Project to understand, at the DNA level, the effects of energy production on human health. The HGP's innovative operational model proved highly successful, and benefits far exceeded the original goal. Today, DOE is poised to take the next vital steps—translating the genetic code in DNA into a new understanding of how life works and applying those biological processes to serve its challenging missions. Effective use of microbial and other biological systems and components will generate new biotechnological industries involving fuels, biochemical processing, nanomaterials, and broader environmental and biomedical applications.

The Office of Science has the capabilities and institutional traditions to bring the biological, physical, and computing sciences together at the scale and complexity required for success. Its academic affiliations, national laboratories, and other resources include major facilities for DNA sequencing and molecular-structure characterization, the high-performance computing resources of the Office of Advanced Scientific Computing Research (OASCR), the expertise and infrastructure for technology development, and a tradition of productive multidisciplinary research essential for such an ambitious research program. In the effort to understand biological systems, these strong assets and the GTL program will complement and extend the capabilities and research efforts supported by the National Institutes of Health, National Science Foundation, other agencies and institutions, and industry.

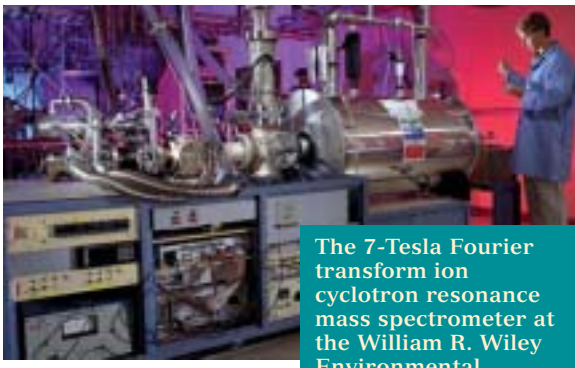
Potential Applications of GTL Science



Learning about the inner workings of microbes and their diverse inventory of molecular machines can lead to discovery of ways to isolate and use these components to develop new, synthetic nanostructures that carry out some of the functions of living cells. In this figure, the enzyme organophosphorus hydrolase (OPH) has been embedded in a synthetic nanomembrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* 124, 11242–43 (2002)]. The OPH transforms toxic substances (purple molecule at left of OPH) to harmless byproducts (yellow and red molecules at right). Applications such as this could enable development of efficient enzyme-based ways to produce energy, remove or inactivate contaminants, and sequester carbon to mitigate global climate change. The knowledge gained from GTL research also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals.

Genomes to Life continues to build its R&D portfolio, having made awards in July 2002 that totaled $103 million for FY 2002–FY 2007. These projects are focusing on isolating and characterizing protein machines, understanding complex biological communities, modeling cellular metabolic and regulatory processes, and modeling carbon sequestration processes in marine microbes.

Projects were chosen to test the concept of systems biology applications and to demonstrate advanced technologies (see picture at right), computation, and potential high-throughput methods in areas having possible impact on DOE missions. These awards represent the culmination of nearly 3 years of planning by the DOE Office of Science and hundreds of scientists at universities, national laboratories, and industry. The microbes studied in the pilot projects, as well as the 2002 awards, have potential for bioremediating metals and radionuclides, degrading organic pollutants, producing energy feedstocks including biomass and hydrogen, sequestering carbon, and demonstrating importance in ocean carbon cycling.



The 7-Tesla Fourier transform ion cyclotron resonance mass spectrometer at the William R. Wiley Environmental Molecular Sciences Laboratory. Mass spectrometry is the most sensitive method for identifying proteins.

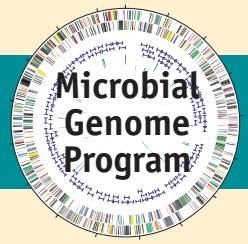
Institutions and Projects Awarded in 2002

- *Oak Ridge National Laboratory:* Developing technologies needed to identify and characterize the complete set of multiprotein complexes within a microbe involved in the carbon cycle (important for carbon sequestration) and another microbe that has the ability to clean up metals in contaminated soil (www.ornl.gov/GenomestoLife/).
- *Lawrence Berkeley National Laboratory:* Developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to environmental conditions found at DOE waste sites (vimss.lbl.gov/).
- *Sandia National Laboratories:* Developing experimental and computational methods to understand the proteins, protein-protein interactions, and gene regulatory networks in a marine microbe that plays a significant role in earth's carbon cycle; important for carbon sequestration (www.genomes-to-life.org/).
- *University of Massachusetts, Amherst:* Developing computational models to predict the activity of natural communities of microbes having potential for uranium bioremediation and production of electricity through their ability to transfer electrons to electrodes (DOEGenomesToLife.org/research/umass.html).
- *Harvard Medical School:* Studying the proteins, protein-protein interactions, gene regulatory networks, and community behavior of microbes active in the carbon cycle (with capabilities relevant to carbon sequestration); important for bioremediation strategies. Developing computational methods to understand the complex biology of these microbes at a systems level (arep.med.harvard.edu/DOEGTL/).

Other Participating Institutions

Argonne National Laboratory
Brigham and Women's Hospital
Diversa Corporation
Los Alamos National Laboratory
Massachusetts General Hospital
Massachusetts Institute of Technology
National Center for Genome Resources
Pacific Northwest National Laboratory
The Institute for Genomic Research
The Molecular Science Institute
University of California (Berkeley, San Diego, Santa Barbara)
University of Illinois
University of Michigan
University of Missouri
University of North Carolina
University of Tennessee (Knoxville, Memphis)
University of Utah
University of Washington

Microbes studied in GTL have had their genetic sequences determined under DOE's Microbial Genome Program.



Direct Web Access

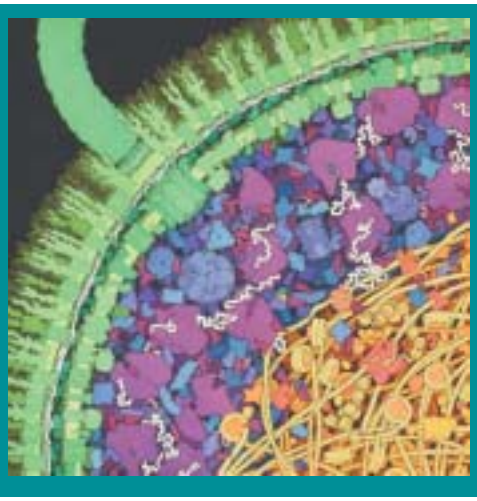
- doegenomestolife.org/pubs.html
- doegenomestolife.org/gallery/images.html
- doegenomestolife.org/research/index.html
- www.ornl.gov/microbialgenomes
- www.ornl.gov/hgmis/education/education.html

FY 2003 Call for Proposals: www.er.doe.gov/production/grants/Fr03-05.html

Integrated, Large-Scale User Facilities: A Plan to Democratize Access to Systems Biology Resources

Analyzing whole microbial systems requires economies of scale. Traditionally, scientists have tried to understand the functions of individual proteins or small groups of proteins. In the new era of systems biology, researchers will study the behavior of the cell's entire working complements of proteins (proteomes), their regulatory pathways, and their interactions as they perform functions. These activities must be carried out on a scale that far exceeds today's capacities.

To meet this challenge, BER and OASCR have planned a set of four core research facilities. Building on each other, these facilities are intricately linked in their long-term goals, targets, technologies, capabilities, and capacities. They will provide scientists with an enduring comprehensive ability to understand and, ultimately, reap enormous benefit from the biochemical functionality of microbial systems. Making the most advanced technologies and computing resources available to scientists in large or small laboratories will democratize access to the tools needed for systems biology. They thus open new avenues of inquiry, fundamentally changing the course of biological research and greatly accelerating the pace of discovery. Hallmarks of these facilities include high-throughput advanced technologies, automation, and tools for data management and analysis, simulation, and an integrated knowledge base.



Crowded cell by D. Goodsell (©1999)

Large-scale facilities are required for identifying, characterizing, and modeling the activity of the tens of thousands of interacting components present at any given moment in a cell.

Large-Scale Facilities Spur Cost, Productivity Improvements

The dramatically increased productivity and reduced costs achieved in the HGP via high-throughput production environments (e.g., the DOE Joint Genome Institute) provide the paradigm for the dedicated industrial-scale facilities envisioned for Genomes to Life.

A Plan for GTL User Facilities

Facility I for Production and Characterization of Proteins would use highly automated processes to mass-produce and characterize proteins directly from microbial genome data and create affinity reagents ("tags") to identify, capture, and monitor the proteins from living systems.

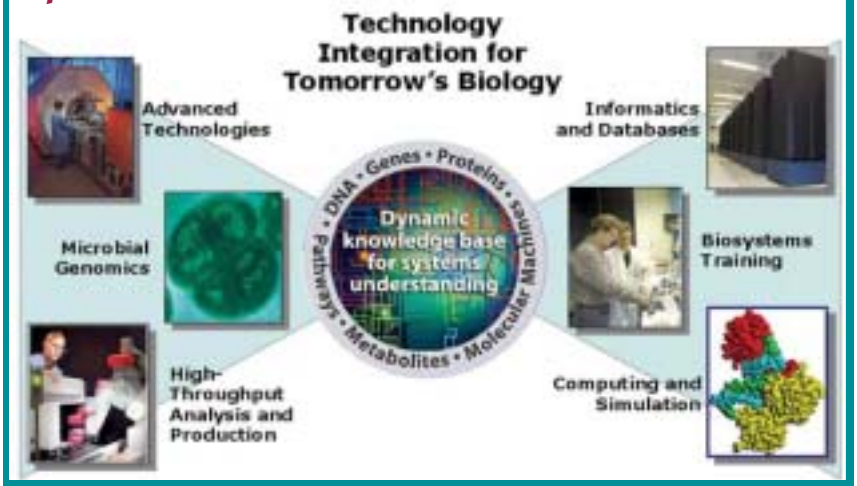
Facility II for Whole Proteome Analysis would characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins.

Facility III for Characterization and Imaging of Molecular Machines would isolate, identify, and characterize thousands of molecular machines from microbes and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells.

Facility IV for Analysis and Modeling of Cellular Systems would combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of spatial and temporal variations in the structures and functions of cellular systems—from individual microbial cells to complex communities and multicellular organisms.

GTL User Facility Hallmarks

Open Access to Data and Facilities



These facilities would serve as focal points for the life sciences research community, providing a national venue to pursue multidisciplinary systems biology and promote cross-disciplinary education.

OSTP, OMB: "National Science Priority"

Achieving a molecular-level understanding of life processes is a national science priority, according to the Office of Science and Technology Policy (OSTP) and Office of Management and Budget (OMB) *FY 2004 Interagency Research and Development Priorities*. The view in this guidance reflects that found throughout much of the biological research community.

AAM: "Develop New Technologies"

Specific recommendations made by the American Academy of Microbiology (AAM) in its 2001 colloquium report, *Microbial Ecology and Genomics: A Crossroads of Opportunity* include the following:

- "Develop new technologies including methods for measuring the activity of microorganisms (at the level of populations and single cells), approaches to cultivating currently uncultivable species, and methods for rapid determination of key physiological traits and activities.
- "Establish mechanisms to encourage the necessary instrument development.
- "Encourage instrumentation development through collaboration with device engineers, chemists, physicists, and computational scientists, since uncovering the diversity and activities of the microbial world is dependent on such advances.
- "Develop technology and analysis capability to study microbial communities and symbioses holistically, measuring system-wide expression patterns (mRNA and protein) and activity measurements at the level of populations and single cells."

BERAC Subcommittee: "Create Unique, High-Throughput Research Facilities"

The BER program of DOE, having played a critical, catalytic role in bringing about the genomic revolution, is now poised to make equally seminal contributions to the next, transforming phase of biology. A subcommittee report approved by the BER Advisory Committee (BERAC) in April 2002 stated: "DOE should now create unique, high-throughput research facilities and resources to translate the new biology, embodied in the Genomes to Life (GTL) program, into a reality for the nation. . . . [GTL] is designed to build on the major accomplishments of the past decade and move from this vision to reality—to a new and comprehensive systems approach from which we will understand the functioning of cells and organisms and their interactions with their environments. Since the science has changed so profoundly, to accomplish these challenging goals in a timely and cost-effective fashion, new facilities and new scientific resources are needed."

GTL Program Development

Genomes to Life is a joint program of the Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research in the Office of Science of the U.S. Department of Energy.

To solicit guidance from the scientific community in the development of the GTL program, in the past 2 years DOE has sponsored 15 workshops, which were attended by scientists from industry, national laboratories, and academia. A strategic plan for developing advanced and high-throughput facilities to serve GTL and the entire community was approved by the BER Advisory Committee (BERAC) in April 2002, and BERAC voiced approval of the subsequent facilities plan in December 2002.

GTL was developed in response to a 1999 charge by the DOE Office of Science to BERAC to define DOE's potential roles in post-HGP science. The resulting report, *Bringing the Genome to Life* (August 2000), set forth recommendations that led to the *Genomes to Life* roadmap (April 2001). Funding for FY 2002 was $21.7 million. The FY 2003 budget for the program proposed in the President's Request to Congress is $42.4 million.

U.S. Department of Energy
Office of Science

Marvin Frazier
Office of Biological and Environmental Research (SC-72)
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

Gary Johnson
Office of Advanced Scientific Computing Research (SC-30)
301/903-5800, Fax: 301/903-7774
gary.johnson@science.doe.gov

Web site for this document:

- **DOEGenomesToLife.org/pubs/overview.pdf**